Preliminary Evaluation of an Aviation Safety Thesaurus' Utility for Enhancing Automated Processing of Incident Reports

Francesca Barrientos, Ph.D.
Joseph Castle
Dawn McIntosh
Ashok Srivastava, Ph.D.

Purpose

The purpose of this document is to present a preliminary evaluation the utility of the FAA Safety Analytics Thesaurus (SAT) in enhancing automated document processing applications under development at NASA Ames Research Center (ARC). Current development efforts at ARC are described, including overviews of the statistical machine learning techniques that have been investigated. An analysis of opportunities for applying thesaurus knowledge to improving algorithm performance is then presented.

Background

The Intelligent Data Mining group at NASA Ames Research Center has been developing machine learning algorithms and software tools to perform text mining and other document processing on the Aviation Safety Reporting System (ASRS) and Aviation Safety Action Program (ASAP) incident report databases. Two different problems are being addressed by this effort. The first is the automated categorization (classification) of incident reports by *event type*. The event types are drawn from the Distributed National ASAP Archive (DNAA) Master List [1] of 31 primary event types, and a report may belong to more than one event type category. The second task is to identify the *contributing factors* associated with each report's events. That is, given a report and its event types, list the contributing factors associated with each event type. The 27 contributing factor labels are also taken from the DNAA Master List.

At present, event types and contributing factors are labeled by hand. Processing ASRS and ASAP incident reports in this way is becoming unfeasible, due to the increasingly large number of reports. Automated categorization of reports has a number of potential advantages over using humans, including scalability and consistency. *Scalability* merely refers to the amount of time and (human) effort required to read through and categorize reports. This scalability issue is especially prominent if the DNAA Master List event type changes and it becomes necessary to recategorize all of the existing reports in the database. Computers can perform this task much faster than humans. *Consistency* can be a problem when manual categorization is performed by different people. With an automated system, inconsistencies between individuals can be eliminated.

To date, our text mining efforts have primarily been applied to the first task, event type categorization [2-4]. We have investigated a number of different of machine learning (ML) approaches, including:

- Support Vector Machines (SVM)

- Naïve Bayes
- Random Forest
- ADAboost

These methods are all statistically based; they build a document classifier from a set of pre-labeled reports based on information about word frequencies. (Concise descriptions of these and other ML techniques appear in [5].)

Our preliminary experiments thus far have produced promising results. In a pilot experiment, an expert was presented with one hundred reports categorized by event type using ML techniques. Each report was labeled with up to five event types ranked in order of confidence. The expert agreed with the top-ranked choice 73% of the time, and with one of the top two choices 86% of the time.

Document Preprocessing

Before classification, text documents are converted into a representation that characterizes their contents in an informative way. In the case of the algorithms listed above, the representation characterizes the frequency and/or importance of each unique *term* that appears in it. The simplest method of generating terms for the document representation is to build a *term-frequency matrix*. Note that this method assumes that individual words are an appropriate semantic unit (lexical semantics) for characterizing the reports.

Other preprocessing steps filter or combine words with the intent of reducing computation and increasing the representation's accuracy. Some amount of natural language processing (NLP) is common to most preprocessing systems. This processing includes acronym expansion, stemming, and combining phrases into a term. Depending on the context, a phrase may have meaning that is lost when only individual words are considered. An example from the aviation domain is "overhead bins." Thesauri are used to combine synonymous words into single term. In our work, we have experimented with the aviation safety-centric PLADS NLP system.

A major difficulty with automated text categorization is applying these algorithms when the number of unique terms is very large. Document sets can have many thousands of unique terms, far more than are manageable using today's computers. An active area of research is developing methods to reduce the number of terms in the document set while minimally affecting accuracy.

One strategy, sometimes referred to as *term selection,* is to select the most informative *subset* of terms. We have experimented with several popular statistical term selection methods. These include information gain (IG), mutual information (MI), and term frequency inverse document frequency (*tf-idf*). (These and other methods are reviewed in [5].)

Some NLP steps, such as combining phrases or synonyms into a single term, reduce the number of terms. Common sense also suggests that applying these methods, especially when combined with domain knowledge, should increase classification accuracy. On the down side, NLP is very expensive computationally, which may outweigh the benefits it

confers. To determine the utility of NLP for our text classification task, we applied ML techniques to both raw and PLADS-preprocessed text. Our initial findings are that NLP preprocessing only marginally improved overall categorization accuracy.

Analysis

The effectiveness of our text mining systems has been improved mainly through optimizing parameters on our machine learning models. As described above, we have used the PLADS NLP preprocessing system to incorporate domain knowledge into our models. PLADS performs the elementary NLP processing that the Safety Analytics Thesaurus (SAT) was designed for, such as stemming, linking synonymous and related terms, and normalizing spelling. Since PLADS only minimally improved performance, trivial preprocessing using the SAT is unlikely to lead to further improvements.

Our statistical techniques perform well in overall categorization, but there are specific cases where miscategorization is more frequent. It is possible that the SAT maybe be useful for handcrafting rules relating to these special cases. We have analyzed each case to determine its possible causes and solutions. Where we feel a thesaurus would contribute to the solution, we have described how it could be applied. Applicability is highly dependent on the thesaurus's topic coverage. The version of the SAT that we used in this analysis is based on ten safety topics.

In one case, our systems have difficulty with accurate *categorization* of a specific event type. The most difficult event type for our systems is *Operation in noncompliance – FARs, policy/procedures*. This event type covers violations of regulations, policies, procedures, and other kinds of rules.

Because our techniques learn from examples of reports, problems with this category imply that there is no consistent language in the reports that correlates to this event type. Examination of the secondary event types reveals that the primary event type is very broad, covering regulations about crew, company policies, federal regulations, weather minimums and equipment. Categories that are derived from a set of disjunctions are inherently difficult for machine learning algorithms. In this case, the solution is to decompose the category into a set of subcategories that are easier to learn, and a natural breakdown is the secondary event type level. The SAT, in its current form, cannot help this problem since the terms in the thesaurus do not cover the topic of regulations and compliance.

In other cases there is a kind of *symmetric confusion* where a report is labeled with an event type that appears to be semantically opposite to the true event. The primary examples of confusion are:

- Excursion/Incursion
- Departure Problems/Landing Event
- Departure Problems /Approach-Arrival Problems

Symmetric confusion implies that reports belonging to one event type have very similar language to the event type with which they are confused.

The pair *Incursion/Excursion* would appear to be difficult to separate since reports relating to either of these event types will use language describing taxiways, runways, and other ground surfaces. Examination of the DNAA Master list shows that *incursion* and *excursion* have virtually identical secondary event types. The secondary level decomposes into kinds of airport surfaces. Conceptually, there are different ways to separate these categories. In the case of *excursion* the cause of the incident is internal to the aircraft; for *incursion*, the cause is external to the aircraft. These event types also result in different types of hazards. The hazard in excursion is the aircraft leaving its designated or intended location. In incursion, the hazard is loss of separation or potential collision. The SAT covers the topic of incursions but not of excursions, so it could not be applied to this problem.

For the pairs involving departure, landing, and approach, reports tend to have similar language relating to air traffic control, clearances, and navigation (e.g., intersections). Conceptually, *Landing Event* and *Approach-Arrival Problems* can be differentiated from *Departure Problems* by the topic of misconfiguration of the aircraft. Even so, there is significant overlap to these concepts because they all relate to flight phases. The SAT does include terms and relationships for traffic control, clearances, and navigation, but these are not related to problems associated with these specific flight phases. It is unclear if the SAT could be applied to this problem.

The last type of case is *similarity confusion* where a report is miscategorized with an event that is closely related to the true categorization. Our confusion cases are:

- Takeoff Deviations/Departure Problems
- Traffic Proximity Event/Airspace Deviation.

For these pairs, the reports have similar language—relating, again, to air traffic control, clearances, and navigation. Traffic proximity can be differentiated by references to the Traffic Collision Avoidance System (TCAS). Our initial analysis does not a include enough examples of these confusion types to make recommendations for category differentiation.

Conclusions

The SAT in its current form is difficult to apply to general classification of incident safety reports since its ten safety topics lack coverage of the DNAA Master List of event types. Incompleteness is a common limitation of thesauri. As the thesaurus expands to cover more topics, this may become less of a problem. Further, results from developing text mining systems such as ours should influence thesaurus development.

Even if there were specific categories where applying the thesaurus would result in improved accuracy, the benefits must be weighed against the effort required to apply domain knowledge. The following factors should be included in any assessment of the manual effort required to use the thesaurus:

- Effort to analyze applicability of thesaurus
- Effort for subject matter experts to develop rules
- Effort to hand-code rules into the classifier

- Length of time between updates to categories (when rules would need to be re-coded and new rules added)
- Time to test rules (added development time)

Even given these disadvantages, we should keep track of the evolution of the SAT as it is expanded to include new safety topics, since this may improve its applicability. Furthermore, with improvements in processor speed, memory, and parallel algorithms, NLP may become less expensive.

Acknowledgements

The material is based upon work supported by NASA under award NCC2-1426.

References

1. Distributed National ASAP Archive Master List Dictionary.

2. A. N. Srivastava, R. Akella, et. al., "Enabling the Discovery of Recurring Anomalies in Aerospace System Problem Reports using High-Dimensional Clustering Techniques," accepted for publication in the 2006 Proceedings of the IEEE Aerospace Conference.

3. A.N. Srivastava, D. McIntosh, J.P. Castle, "Automatic Discovery of Anomalies Reported in Aerospace System-Health and Safety Documents," submitted to AIAAInfotech, 2007.

4. A. N. Srivastava, B. Zane-Ulman, "Discovering Hidden Anomalies in Text Reports Regarding Complex Space Systems", IEEE Aerospace Conference, Big Sky, MT, 2005.

5. Sebastiani, Fabrizio, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1-47.